Thermodynamic Rule Determining the Biological DNA Information Capacity

A. Widom and J. Swain
Physics Department, Northeastern University, Boston MA USA

Y.N. Srivastava
Physics Department & INFN, University of Perugia, Perugia IT

S. Sivasubramanian
Center for High-Rate Nanomanufacturing, Northeastern University, Boston, MA USA

V.I. Valenzi
Centro Studi di Biometeorologia Onlus Roma/Lugano Via Besso 59, Lugano CH

A rigorous thermodynamic expression is derived for the total biological information capacity per unit length of a DNA molecule. The total information includes the usual four letter coding sequence information plus that excess information coding often erroneously referred to as "junk". We conclude that the currently understood human DNA code is about a hundred megabyte program written on a molecule with about a ten gigabyte memory. By far, most of the programing code is not presently understood.

PACS numbers: 82.39.Pj, 87.14.gk, 87.14.gn

I. INTRODUCTION

The information capacity $\mathcal{H}$ in a human DNA molecule of length $L \sim 3$ meter arising from the conventional four letter sequence code has been estimated to be[1]

$$\mathcal{H}_{\rm code} \sim 10^9 \text{ bit.} \tag{1}$$

This estimate is approximately correct for humans, apes and perhaps snails. While the authors might admit some similarities with the apes, they would object to being called similar to snails. In defense of claiming that humans are higher on the evolutionary scale, one might include the so-called "junk" DNA residing in the chain. The estimates are

$$\mathcal{H}_{\rm code+junk} \sim 10^{11} \text{ bit.} \tag{2}$$

In terms of the total information capacity, humans do indeed appear to be on a higher evolutionary scale than (say) snails.

There has been considerable recent interest in the nature of "junk" DNA sequences[2–5] and in particular the role that they play in the evolutionary process. Our purpose is to derive a thermodynamic expression for the information which resides in DNA. In the thermodynamic limit, the information capacity per unit length

$$\eta = \lim_{L\to\infty} \frac{\mathcal{H}}{L} \tag{3}$$

has a thermodynamic description which is both mathematically rigorous and yet is experimentally measurable. The rule may be stated in terms of the DNA chain tension τ as a function of temperature T and the chemical potentials $(\mu_1, \mu_2, \ldots, \mu_c)$ of the molecules which make up the DNA chain; i.e.

$$\tau = \tau(T, \mu_1, \mu_2, \ldots, \mu_c) \tag{4}$$

Our central result concerns a precise expression for η;
Theorem: *The information capacity per unit length of a DNA chain is given by the thermodynamic expression*

$$\eta = -\left[\frac{1}{k_B \ln 2}\right]\left(\frac{\partial \tau}{\partial T}\right)_{\mu_1,\mu_2,\ldots,\mu_c}. \tag{5}$$

wherein k_B is Boltzmann's constant.

The rigorous proof of the theorem will be given in Sec.III. The only assumptions of the proof reside in the first and second laws of statistical thermodynamics. Otherwise, the theorem is completely model independent. The importance of a force determination of information capacity is that in the laboratory tweezer measurements, either optical[6] or magnetic[7], uniquely determine the DNA chain tension $\tau(T, \mu_1, \mu_2, \ldots, \mu_c)$.

In Sec.II the relationship between thermodynamic entropy $\mathcal{S}$ and information $\mathcal{H}$ is reviewed. The statistical thermodynamics of long chain molecules is explored in Sec.III and the proof of the central theorem is provided. An order of magnitude statement for η in the human genome is discussed in the concluding Sec.IV. As one moves up the evolutionary scale, the total information capacity in the DNA molecule appears to increase.

II. ENTROPY AND INFORMATION

The connection between entropy and information in statistical thermodynamics is well understood[8]. The number of microscopic states Ω of a macroscopic system may be written as

$$\Omega = e^{\mathcal{S}/k_B} = 2^{\mathcal{H}}, \tag{6}$$

wherein the thermodynamic entropy $\mathcal{S}$ is determined by

$$\mathcal{S} = k_B \ln \Omega \quad \text{where} \quad k_B \approx 1.38065 \times 10^{-16} \ \frac{\rm erg}{^{\rm o}{\rm K}}. \tag{7}$$

The information capacity measured in bits[9] is thereby

$$\mathcal{H} = \lg\,\Omega = \frac{\ln \Omega}{\ln 2} = \frac{\mathcal{S}}{k_B \ln 2}\ . \tag{8}$$

where $\ln \equiv \log_{\rm e}$ and $\lg \equiv \log_2$. A discussion of the thermodynamic entropy of a DNA chain follows.

III. STATISTICAL THERMODYNAMICS

If $\mathcal{E}$ denotes the energy of a molecular chain of length L and molecular composition numbers $\mathcal{N}_1, \mathcal{N}_2, \ldots, \mathcal{N}_c$, then the first and second thermodynamic laws for quasi static processes read

$$d\mathcal{E} = Td\mathcal{S} + \tau dL + \sum_{j=1}^{c} \mu_j d\mathcal{N}_j\ . \tag{9}$$

The DNA chain quantities $(\mathcal{E}, \mathcal{S}, L, \mathcal{N}_1, \mathcal{N}_2, \ldots, \mathcal{N}_c)$ are all extensive.

Employing extensive scaling

$$\lambda\mathcal{E} = \mathcal{E}(\lambda\mathcal{S}, \lambda L, \lambda\mathcal{N}_1, \lambda\mathcal{N}_2, \ldots, \lambda\mathcal{N}_c), \tag{10}$$

one finds the Euler equation

$$\mathcal{E} = \mathcal{S}\frac{\partial \mathcal{E}}{\partial \mathcal{S}} + L\frac{\partial \mathcal{E}}{\partial L} + \sum_{j=1}^{c} \mathcal{N}_j \frac{\partial \mathcal{E}}{\partial \mathcal{N}_j}\ . \tag{11}$$

Eqs.(9) and (11) imply

$$\mathcal{E} = T\mathcal{S} + \tau L + \sum_{j=1}^{c} \mu_j \mathcal{N}_j\ . \tag{12}$$

Taking the differential of Eq.(12) and comparing the result to Eqs.(9) yields

$$\mathcal{S}dT + Ld\tau + \sum_{j=1}^{c} \mathcal{N}_j d\mu_j = 0. \tag{13}$$

Defining the entropy per unit length σ and the molecular densities per unit length $(\Gamma_1, \Gamma_2, \ldots, \Gamma_c)$ by

$$\sigma = \lim_{L\to\infty} \frac{\mathcal{S}}{L} \quad \text{and} \quad \Gamma_j = \lim_{L\to\infty} \frac{\mathcal{N}_j}{L} \tag{14}$$

together with Eq.(13) yields

$$d\tau = -\sigma dT - \sum_{j=1}^{c} \Gamma_j d\mu_j\ . \tag{15}$$

The entropy per unit length is thereby

$$\sigma = -\left(\frac{\partial \tau}{\partial T}\right)_{\mu_1, \mu_2, \ldots, \mu_c} = k_B \eta \ln 2 \tag{16}$$

allowing for the verification of our central theorem. Eqs.(3), (8), (14) and (16) yield the required proof of Eq.(5).

IV. CONCLUSION

In order to apply the theorem Eq.(5), one has to fix the molecular chemical potentials. These chemical potentials depend on the solution properties of the environment in which the DNA molecule resides. Changing these environmental parameters also changes the information capacity per unit length of the DNA molecule. We here stress that the thermodynamic rule includes the total information capacity of all the possible biophysical forms, e.g. four letter coding, "junk" DNA insertions as well as semiconducting electrons existing in ordered water shells coating the DNA chain. Typical values for the human genome are of order $\eta \sim 10$ byte/nanometer. This is completely consistent with Eq.(2). The total information in a DNA molecule gets larger as one moves up the evolutionary scale. We conclude that the currently understood DNA code is about a 100 megabyte program written on a molecule with about 10 gigabyte of memory capacity. Most of the programing code is beyond our understanding.

[1] J.K. Percus, "*Mathematics of the Genome Analysis*", Cambridge Studies in Mathematical Biology, Cambridge University Press, Cambridge (2002).

[2] A.T. Willingham and T.R. Gingeras, *Cell* **125**, 1215 (2006).

[3] D.E. Riley and J.N. Krieger, *Biochemical and Biophysical Research Communications* **298**, 581 (2002).

[4] W. Makalowski, *Science* **300**, 1246 (2003).

[5] T.R. Gregory and P.D.N. Hebert, *Genome Res.* **9**, 317 (1999).

[6] C. Bustamante, Z. Bryant and S.B. Smith, *Nature* **421**, 423 (2003).

[7] S.H. Leuba, M.A. Karymov, M. Tomschik, R. Ramjit, P. Smith and J. Zlatanova, *PNAS* **100**, 495 (2003).

[8] L.D. Landau and E.M. Lifshitz, "*Statistical Physics*" Third Edition Part 1, Butterworth-Heinemann, Oxford (1999).

[9] C. Adami, *Physics of Life Reviews* **1**, 3 (2004).